

ON BERNSTEIN - VON MISES THEOREM AND SURVIVAL ANALYSIS

Jana Timková

timkova@karlin.mff.cuni.cz

Department of Statistics, Charles University, Prague
Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Prague

Summary: Although not well-known, the Bernstein-von Mises theorem (BvM) is a so-called bridge between bayesian and frequentist asymptotics. Basically, it states that under mild conditions the posterior distribution of the model parameter centered at the maximum likelohood estimator (MLE) is asymptotically equivalent to the sampling distribution of the MLE. This is a powerful tool especially when the classical asymptotics is tedious or impossible to conduct while bayesian asymptotic properties can be obtianed via MCMC. However, in semiparametric setting with presence of infinite-dimensional parameters, as is e.g. Cox model for survival data, the results regarding BvM are more difficult to establish but still not impossible. The proposed poster gives short overview of BvM results found in the survival analysis context.

Bernstein - von Mises theorem: the parametric case

Let $X_i, i = 1, .., n$ be independent identically distributed random variables distributed according to $P_\theta, \theta \in \Theta \subset \mathbb{R}^p$, Θ open. Let $f(x, \theta)$ be a probability density of P_θ with respect to Lebesgue measure. Suppose that $\pi(\theta)$ is an a priori probability density on Θ which is continuous and positive in an open neighbourhood of the true parameter θ_0. Suppose that $\partial \log f(x,\theta)/\partial\theta$ and $\partial^2 \log f(x,\theta)/\partial\theta^2$ exist and are continuous in θ. Further, suppose that $I(\theta) = -\,\mathrm{E}\left[\partial^2 \log f(x,\theta)/\partial\theta^2\right]$ is continuous, with $0 \le I(\theta) \le \infty$.

Let θ_n be **a maximum-likelihood estimator of** θ based on $X_1, .., X_n$.

Theorem 1 (Frequentist asymptotics for MLE of $\boldsymbol{\theta}$, [1]) *Under certain regularity conditions*

$$\sqrt{n}(\hat\theta_n - \theta_0) \xrightarrow{\mathscr{D}} \mathscr{N}(0, I(\theta_0)^{-1}).$$

Let $\pi(\theta)$ denote the prior probability density of θ and $\pi_n(\theta|x_1, .., x_n)$ denote the posterior density of θ given the observation $x_1, .., x_n$. Further denote $\pi_n^\star(t|x_1, .., x_n) = n^{-1/2}\pi_n(\hat\theta_n + t/\sqrt{n}|x_1, .., x_n)$. Then $\pi_n^\star(t|x_1, .., x_n)$ is **the a posteriori density of the rescaled parameter** $t = \sqrt{n}(\theta - \hat\theta_n)$. Let $\Pi^\star(dt|x_1, .., x_n)$ be the probability measure with density $\pi^\star(t|x_1, .., x_n)$ and P_θ^n be the joint distribution of $X_1, .., X_n$.

Theorem 2 (Parametric Bernstein - von Mises, [5]) *Let $\{P_\theta, \theta \in \Theta\}$ be differentiable in quadratic mean at θ_0 with nonsingular Fisher information I_{θ_0}, and suppose that for every sequence of balls $(K_n)_{n\ge1} \subset \mathbb{R}^p$ with radii $M_n \to \infty$, we have*

$$\Pi^\star(K_n|X_1, .., X_n) \xrightarrow{P_{\theta_0}^n} 1.$$

Then the posterior distribution $\Pi^\star(dt|X_1, .., X_n)$ of the scaled parameter $t = \sqrt{n}(\theta - \hat\theta)$, given $X_1, .., X_n$, satisfies

$$\sup_{B\in\mathscr{B}^p} |\Pi^\star(B|X_1, .., X_n) - \Phi(B)| \xrightarrow{P_{\theta_0}^n} 0,$$

where Φ is the probability measure of normal distribution with mean zero and variance $I(\theta_0)^{-1}$ and $\mathscr{B}^p$ denotes the set of all Borel subsets on $\mathbb{R}^p$.

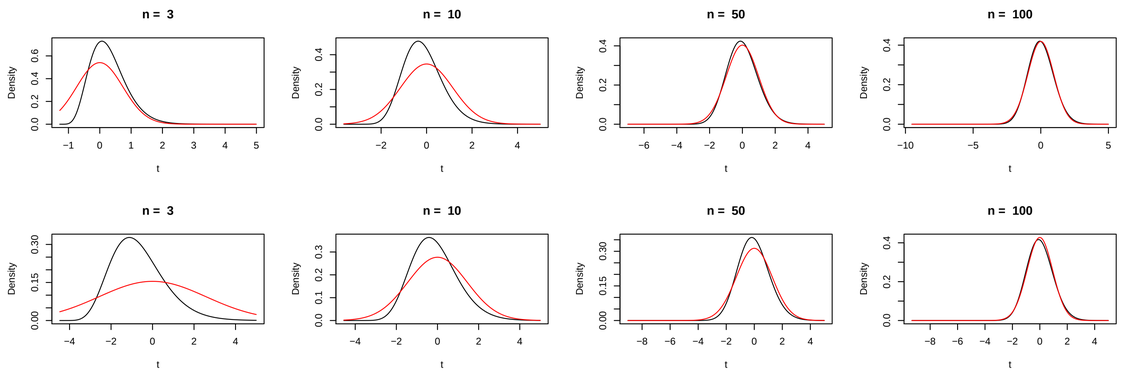

Figure 1. Example: Generate $x_1, .., x_n$, an n-sample from exponential distribution with the true parameter $\lambda_0 = 1$ (first row) and $\lambda_0 = 3$ (second row). ML estimate of λ is the mean $\overline{x} = \sum_i x_i/n$. Take a family of cunjugate priors $Gamma(a, b)$ for λ. The posterior density of $\sqrt{n}(\lambda - \overline{x})$ and corresponding normal density are in black and red, respectively. The size of sample is $n = 3, 10, 50$ and 100 from left to right.

Bernstein - von Mises theorem in semiparametric setting: Cox model

Let $X_i, i = 1, .., n$, be survival times and alongside with each X_i let us observe a set of covariates $\boldsymbol{Z}_i \in \mathbb{R}^p$. Let $F_i(\cdot)$ be the distribution function associated with X_i. As usual in applications, the survival times are assumed to be **right-censored**, that is, the actual observed dataset is a set of triplets $(T_i, \delta_i, \boldsymbol{Z}_i, i = 1, .., n)$ where $T_i = \min(X_i, C_i)$, $\delta_i = I(T_i = X_i)$ and C_i is censoring random variable independent on X_i.

The traditional approach to specify **Cox model** is via particular form of the hazard rate which is assumed to satisfy

$$\Lambda_i(t) = \Lambda(t, \boldsymbol{Z}_i) = \int_0^t \exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}_i\} d\Lambda(s), \quad i = 1, .., n,$$

where $\boldsymbol{\beta}$ is unknown p-dimensional regression parameter and Λ is an unknown cumulative hazard rate of a survival time with the covariate being equal to 0. After minor calculations we can reach an alternative formulation

$$1 - F_i(t) = (1 - F(t))^{\exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}_i\}}, \quad i = 1, .., n,$$

where F is an unknown distribution for an individual with $\boldsymbol{Z} = 0$.

Remark: There is a one-to-one relation between F_i and Λ_i, precisely $\Lambda_i(t) = \int_0^t \frac{\Delta F_i(s)}{1 - F_i(s_-)} ds$.

Remark: With Λ being functional (so an infinitely-dimensional) parameter and $\boldsymbol{\beta}$ finite-dimensional parameter inference on Cox model falls among **the semiparametric problems**.

Traditional approach to estimate the unknown parameters β and Λ ...

.. is based on **partial likelihood** theory. Let $\boldsymbol{\beta}_0$ and Λ_0 be the true parameters and $\tau = \max\{T_i, i = 1, ..n\}$. The estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is defined as a solution to the vector equation $U(\boldsymbol{\beta}) = 0$ where

$$U(\boldsymbol{\beta}) = \sum_{i=1}^n \left\{ \boldsymbol{Z}_i - \frac{\sum_{j:T_j \ge T_i} \boldsymbol{Z}_j \exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}_j\}}{\sum_{k:T_k \ge T_i} \exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}_k\}} \right\}.$$

The cumulative baseline hazard function $\Lambda(t)$ is estimated using the Breslow estimator

$$\hat\Lambda(t) = \sum_{i:T_i \le t} \frac{\delta_i}{\sum_{j\in R_i} \exp\{\hat{\boldsymbol{\beta}}^\top \boldsymbol{Z}_j\}}.$$

We need to introduce some neccesary **notation** before stating next theorem:

$$\Sigma(\boldsymbol{\beta}_0, t) = \int_0^t \left\{ \frac{s^{(2)}(\boldsymbol{\beta}_0, s)}{s^{(0)}(\boldsymbol{\beta}_0, s)} - \mathbf{e}(\boldsymbol{\beta}_0, s)\mathbf{e}(\boldsymbol{\beta}_0, s)^\top \right\} s^{(0)}(\boldsymbol{\beta}_0, s) d\Lambda_0(s)$$

where

$$\mathbf{e}(\boldsymbol{\beta}_0, s) = \frac{s^{(1)}(\boldsymbol{\beta}_0, s)}{s^{(0)}(\boldsymbol{\beta}_0, s)} \quad \text{and} \quad s^{(l)}(\boldsymbol{\beta}, s) = \lim_{n\to\infty} \frac{1}{n} \sum_{i:T_i \ge t} \boldsymbol{Z}_i^{\otimes l} \exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}_i\}, \quad l = 0, 1, 2.$$

In next we will also use the processes defined as

$$V_0(t) = \int_0^t \frac{d\Lambda_0(s)}{s^{(0)}(\boldsymbol{\beta}_0, s)} \quad \text{and} \quad E_0(t) = \int_0^t \mathbf{e}(\boldsymbol{\beta}_0, t) d\Lambda_0(s).$$

Theorem 3 (Frequentist asymptotics for Cox model, [2]) *Let the conditions A-D in [2] be fulfilled. Then the following is true:*

1. $$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{\mathscr{D}} \mathscr{N}(0, \Sigma(\boldsymbol{\beta}_0, \tau)^{-1})$$

2. $$\mathscr{L}(\sqrt{n}(\hat\Lambda(\cdot) - \Lambda_0(\cdot))|\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = x) \xrightarrow{\mathscr{D}} W(V_0(\cdot) - xE_0(\cdot))$$

on the space of functions continuous to the right and with limits to the left, $D[0, \tau]$. W denotes the standard Brownian motion.

Bayesian: Process neutral to the right as a prior for F

A prior process on the c.d.f. F is a **process neutral to the right** if corresponding Λ is a positive nondecreasing independent increment process (a nonstationary subordinator in the language of Lévy processes, further NII) such that $\Lambda(0) = 0, 0 \le \Delta\Lambda(t) \le 1$ for all t w.p. 1 and either $\Delta\Lambda(t) = 1$ for some $t > 0$ or $\lim_{t\to\infty} \Lambda(t) = \infty$ w.p. 1.

The **Lévy measure** ν of an NII process is defined

$$\nu([0, t] \times B) = \mathrm{E}\left(\sum_{s\in[0,t]} I(\Delta\Lambda_0(s)) \in B \setminus \{0\} \right)$$

where $t \ge 0$, B is a Borel subset of $[0, 1]$. Here we **let the baseline c.d.f. F be, a priori, a process neutral to the right and let us assume that the corresponding Λ is an NII process with the Lévy measure**

$$\nu(dt, dx) = \frac{1}{x} g_t(x)\phi(t)\, dx\, dt, \quad t \ge 0, x \in [0, 1],$$

where $\int_0^1 g_t(x)dx = 1$, $\forall t$, and ϕ is bounded and positive on $[0, \tau]$. And **let $\pi(\boldsymbol{\beta})$ be prior distribution for $\boldsymbol{\beta}$** which is continuous at $\boldsymbol{\beta}_0$ with $\pi(\boldsymbol{\beta}_0) > 0$, where $\boldsymbol{\beta}_0$ is true value of $\boldsymbol{\beta}$.

Let us suppose that usual conditions regarding the regularity of the model are met (see (A1)-(A5) in [3]). Further, **for bayesian asymptotics consider two important conditions**:

(a) There exists $\zeta > 0$ such that

$$\sup_{t\in[0,\tau],\ x\in[0,1]} g_t(x)(1 - x)^{1-\zeta} < \infty.$$

(b) There exists a function $k(t)$ on $[0, \tau]$ such that $0 < \inf_{t\in[0,\tau]} k(t) \le \sup_{t\in[0,\tau]} k(t) < \infty$ and, for some $\alpha \le 1/2$ and $\epsilon > 0$,

$$\sup_{t\in[0,\tau],\ h\in[0,\epsilon]} \left| \frac{g_t(h) - k(t)}{h^\alpha} \right| < \infty.$$

Theorem 4 (Bernstein - von Mises for Cox model, [4]) *Under conditions (a) and (b), the following holds:*

1. $$\lim_{n\to\infty} \int_{\mathbb{R}^p} |f_n(x) - \phi(x)| dx = 0$$

with probability 1, where f_n is the marginal posterior density of $x = \sqrt{n}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$ and ϕ is the normal density with mean 0 and variance $\Sigma(\boldsymbol{\beta}_0, \tau)^{-1}$.

2. $$\mathscr{L}(\sqrt{n}(\Lambda(\cdot) - \hat\Lambda(\cdot)|\sqrt{n}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) = x, (T_i, \boldsymbol{Z}_i, \delta_i)_{i=1}^n) \xrightarrow{\mathscr{D}} W(V_0(\cdot) - xE_0(\cdot)) \qquad (1)$$

on the space of functions continuous to the right and with limits to the left, $D[0, \tau]$, with probability 1, as $n \to \infty$. W denotes the standard Brownian motion.

As a direct result of Theorem 3 we have the convergence of the joint posterior distribution

$$\mathscr{L}(\sqrt{n}(\Lambda(\cdot) - \hat\Lambda(\cdot), \boldsymbol{\beta} - \hat{\boldsymbol{\beta}})|(T_i, \boldsymbol{Z}_i, \delta_i)_{i=1}^n) \xrightarrow{\mathscr{D}} (W(V_0(\cdot) - XE_0(\cdot)), X)$$

with probability 1, as $n \to \infty$ on $D[0, \tau]$. X represents p-dimensional multivariate normal distribution with mean 0 and variance $\Sigma(\boldsymbol{\beta}_0, \tau)^{-1}$. Now let A be an arbitrary **Hadamard-differentiable functional of model parameters** $(\Lambda, \boldsymbol{\beta})$. Then similar result can be obtained for $\sqrt{n}(A(\boldsymbol{\beta}, \Lambda) - A(\hat{\boldsymbol{\beta}}, \hat\Lambda))$ by simply applying the functional delta method (e.g. [5], Section 20.2).

Remark: Useful examples of Hadamard-differentiable functionals:

- survival function $S(t) = \prod_{s\le t}\{1 - d\Lambda(s)\}$
- median residual life η_{t_0} so that $S(\eta_{t_0})/S(t_0) = 0.5$, for $t_0 \in (0, \tau)$.

Example

For illustration we picked the simplest case of noncensored data without covariates. Then the only unknown parameter is cumulative hazard rate $\Lambda(\cdot)$ and the asymptotic distribution of (1) in Theorem 3 simplifies into $W(U_0(\cdot))$ where $U_0(t) = \int_0^t d\Lambda_0/Pr(T \ge t)$.

As a prior process we used **compound Poisson process** with Lévy measure $\nu(dt, dx) = c\sigma(x)\, dx\, dt$ where $\sigma(\cdot)$ is the jump size distribution density. For $\sigma(\cdot)$ we used Beta distribution with parameters $a = 0.1$ and $b = 0.2$. The simulation of posterior density was done using Markov Chain Monte Carlo methods. The number of iterations was 10000, first 5000 were discarded as burn-in. Number of simulated observed failures was 25.

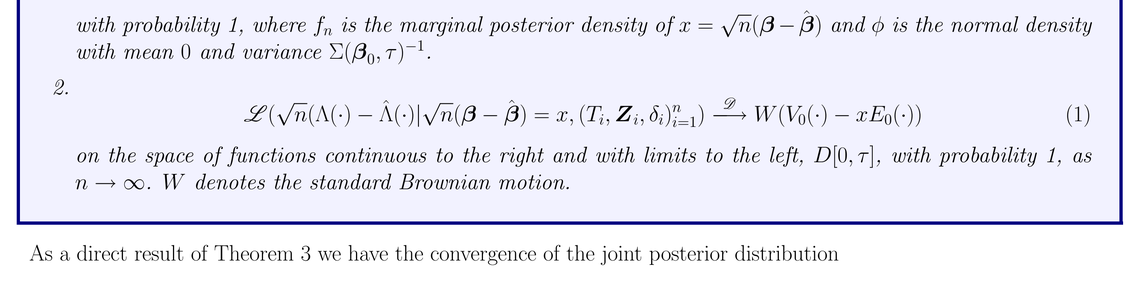

Figure 2. From left to right: **1.** Several iterations from the posterior sample of the cumulative hazard rate $\Lambda(\cdot)$ with frequentists estimator in red color alongside; dots in the bottom are observed data **2.** The posterior median and 95% pointwise credibility band for the cumulative hazard rate with frequentists estimator in red color alongside; dots in the bottom are observed data **3.** Several iterations from the transformed cumulative hazard rate

$$\frac{1}{\hat U(\tau)^{1/2}} \sqrt{n} \left(\Lambda(\cdot) - \hat\Lambda(\cdot) \right)$$

which is asymptotically distributed as the standard Brownian motion and $\hat U(\cdot)$ is a consistent estimator of $U_0(\cdot)$ **4.** The posterior median and 95% pointwise credibility bands for the transformed cumulative hazard rate alongside with the Kolmogorov-Smirnov type bands in red color.

Acknowledgement. The poster was supported by grants GA ČR 201/05/H007 and by GA AV IAA101120604.

References.

[1] Anděl, J. (2002): *Základy matematické statistiky.* Preprint.

[2] Andersen, P. K., Gill R. D. (1982): Cox's regression model for counting processes: A large sample study, *Ann. Statist.* 10, pp. 1100 – 1120.

[3] De Blasi, P., Hjort, N. L. (2009): The Bernstein-von Mises theorem in semiparametric competing risks models, *J. Stat. Plan. and Infer.* Vol. 34, No. 4, pp. 1678 – 1700.

[4] Kim, Y.(2006): The Bernstein-von Mises theorem for the proportional hazard model, *Ann. Statist.* 34, no. 4, pp. 1678 – 1700.

[5] Vaart, A. W. van der (2000): *Asymptotic statistics (Cambridge Series in Statistical and Probabilistic Mathematics)* . Cambridge University Press.